THERMALLY-AWARE THROTTLING IN A THREE-DIMENSIONAL PROCESSOR STACK

BACKGROUND

[0001] Field of the Disclosure

[0002] The present disclosure relates generally to processor systems and, more particularly, to three-dimensional processor stacks.

[0003] Description of the Related Art

[0004] Conventional processing systems are based on two-dimensional (2-D) structures such as a system-on-a-chip (SoC), which may include a variety of components of different sizes and processing capabilities. For example, a heterogeneous SoC may include a combination of processor cores such as one or more central processing units (CPUs), one or more graphics processing units (GPUs), or one or more specialized hardware accelerator processors. A higher level of integration can be achieved by implementing the processing system as a three-dimensional (3-D) structure formed by stacking and interconnecting multiple silicon layers that each include one or more processor cores. The stacked silicon layers in the 3-D processor stack are separated by distances of tens to hundreds of microns and exhibit a high degree of thermal coupling. Thus, heat generated in one or more processor cores of one silicon layer can raise the temperature of the processor cores in the other silicon layers in the 3-D processor stack.

BRIEF DESCRIPTION OF THE DRAWINGS

[0005] The present disclosure may be better understood, and its numerous features and advantages made apparent to those skilled in the art by referencing the accompanying drawings. The use of the same reference symbols in different drawings indicates similar or identical items.

[0006] FIG. **1** is a block diagram of a processing device that includes a 3-D processor stack in accordance with some embodiments.

[0007] FIG. **2** is a contour plot of a thermal density map for a processing device such as the processing device shown in FIG. **1** according to some embodiments.

[0008] FIG. **3** is a diagram of temperatures at different locations in a 3-D processor stack as a function of time according to some embodiments.

[0009] FIG. **4** is a flow diagram of a method of generating a thermal sensitivity map according to some embodiments.

[0010] FIG. **5** is a flow diagram of a method for selectively throttling processor cores in layers of a 3-D processor stack according to some embodiments.

[0011] FIG. **6** is a diagram of a data structure that includes information indicating thermal couplings between locations in a processing system according to some embodiments.

DETAILED DESCRIPTION

[0012] Thermal management techniques that are appropriate for 2-D structures often are less effective or even counterproductive when applied to 3-D structures. For example, static assignment of threads to different processor cores prior to execution of the threads cannot, by definition, be used to address all possible thermal emergencies at run-time. As used herein, the term "thermal emergency" refers to the temperature of a processor core or other entity in a processing system exceeding a threshold temperature that indicates a potential for damage due to overheating. For another example, dynamically migrating threads from high temperature processor cores to low temperature processor cores incurs significant performance overhead and may not effectively reduce the temperature in the 3-D structure due to the high degree of thermal coupling between the layers. For yet another example, throttling threads based on their power density or power consumption may incur a significant performance loss because high power density (or consumption) threads are typically performance critical threads.

[0013] Thermal emergencies in a 3-D processor stack can be avoided or mitigated at run-time by selectively throttling one or more of a plurality of processor cores implemented in a plurality of layers of the 3-D processor stack based on values of thermal couplings between the plurality of processor cores in the plurality of layers and based on measures of criticality of threads executing on the plurality of processor cores. In some embodiments, the values of the thermal couplings indicate temperature changes in each of the plurality of layers (or each of the plurality of processor cores) as a function of temperature changes in each of the other layers (or each of the other processor cores). For example, the values of the thermal couplings may indicate a level or degree of thermal coupling between different locations. The values of the thermal couplings may also include temporal information related to the thermal coupling values. For example, the temporal information may indicate latencies between temperature changes in each of the plurality of layers or processor cores. In some embodiments, the values of the thermal couplings indicate coarse levels of the thermal couplings, such as a low level of thermal coupling to indicate that a temperature change in a layer or processor core has a small thermal impact on the temperature in another layer or processor core, a medium level of thermal coupling to indicate that a temperature change in a layer or processor core has a moderate thermal impact on the temperature in another layer or processor core, and a high level of thermal coupling to indicate that a temperature change in a layer or processor core has a large thermal impact on the temperature in another layer or processor core.

[0014] Measures of the criticality of the threads may include indicators of criticality provided by an operating system or values of hardware event counters associated with the threads, such as instruction counts or floating-point operation counts. In some embodiments, selectively throttling one or more of the processor cores includes predicting or detecting a thermal event associated with a critical thread in a first layer (or a first processor core in the first layer) and, in response, throttling a second layer (or a second processor core) that has a strong thermal coupling to the first layer and is executing a non-critical thread. Additional non-critical threads in other layers or cores may also be throttled to provide additional thermal impact and, if throttling non-critical threads does not mitigate the thermal emergency, the critical thread may be throttled.

[0015] FIG. **1** is a block diagram of a processing device **100** that includes a 3-D processor stack in accordance with some embodiments. The 3-D processor stack includes a plurality of layers **150**, **151**, **152** (referred to collectively as "the layers **150-152**") that may be formed on individual substrates or dies. The layers **150-152** are interconnected using any of a variety of interconnect structures, such as pins, balls, traces, wires, interposers, and the like. Although three layers **150-152** are shown in FIG. **1**, some embodiments of the 3-D processor stack may include different